

웨이블릿 변환과 Temporal Fusion Transformer 모델을 활용한 주가지수 예측

최우성¹, 손형오², 류병석³, 김영균⁴
¹강원대학교 행정학전공
²강원대학교 정보통계학전공
³연세대학교 공과대학 화공생명공학과
⁴융합소프트웨어랩

cwsu0313@naver.com, daily6003@gmail.com, bsryu@yonsei.ac.kr, ygkim-2004@hanmail.net

Predicting Stock Price Index using the Wavelet Transform and Temporal Fusion Transformer Model

Useong Choe¹, Hyoengoh Son², Byeongseok Ryu³, YoungGyun Kim⁴

¹Dept. of Public Administration, Kangwon National University

²Dept. of Information and Statistics, Kangwon National University

³Dept. of Chemical & Biomolecular Engineering, Yonsei University

⁴Convergence Software Lab.

요 약

정보기술의 발전으로 여러 분야에 머신러닝 기법이 적용되고 있는 중, 금융·경제 분야는 낮은 신호 대 잡음 비, 머신러닝의 낮은 설명 가능성으로 인하여 머신러닝 활용의 확산이 더디게 일어나고 있다. 이러한 문제점을 해결하고자 웨이블릿 변환을 이용하여 금융데이터의 노이즈를 제거하고 설명 가능성을 높인 딥러닝 모델인 Temporal Fusion Transformer를 활용하여 금융·경제 분야 내 머신러닝 적용의 한계를 극복하는데 기여하고자 한다.

1. 서론

최근 여러 분야에서 기계학습(Machine learning) 기법을 도입하여 새로운 가치를 창출하고 있으나 모델의 복잡성으로 인해 금융경제 분야에서는 기계학습 모형의 도입에 어려움을 겪는다[1]. 예를 들어 금융데이터 중 거시 금융에서의 수년 치 표준 데이터 세트는 수백 개의 월별 관측치에 국한되는 ‘스몰데이터’ 문제를 지니며[2] 시장의 힘이 가격의 예측 가능성을 제거하기 위해 끊임없이 작용하여 신호 대 잡음 비가 낮다는 한계가 있다[3-4]. 또한, 기계학습은 기존 전통적 계량경제학 모델에 비해 설명 가능성이 낮다는 문제가 있다. 기계학습의 예측 결과에 대한 부족한 설명력이나 변수에 대한 분석이 부족한 점은 설명력과 신뢰성이 중요한 금융·경제 분야에 기계학습모델 도입을 저해하는 요소로 작용한다[5]. 이러한 한계에도 불구하고 기계학습 모형은 높은 예측 정확도를 보이는 장점을 가지고 있어, 금융·경제 분야에도 기계학습 모형을 적용하려는 시도가 이루어지고 있다[2]. 본 논문에서는 금융경제 분야의 기계학습 방법의 적용을 저해하는 요소인 노이즈를 줄이고 설명력을 보완하고자 웨이블릿

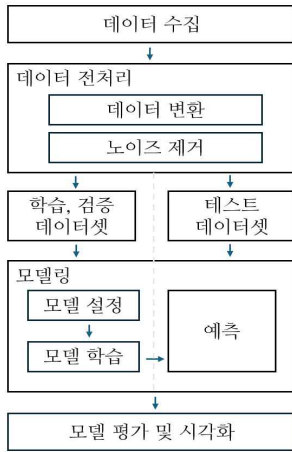
변환(Wavelet Transform)을 통한 디노이징을 금융 시계열 데이터에 적용하고, 어텐션(Attention) 메커니즘으로 해석 가능성을 높인 딥러닝 모델인 Temporal Fusion Transformer 모델을 이용하여 주가지수를 예측하고자 한다.

2. 연구 설계

2.1 Temporal Fusion Transformer

Temporal Fusion Transformer(TFT)는 다중 기간 예측(Multi-horizon forecasting)을 위한 딥러닝 모델이다[6]. TFT는 어텐션 기법을 활용한 구조이며, 여러 규모의 시간적 패턴을 도출하기 위한 셀프 어텐션(Self attention) 층을 가지고 있다[7]. 또한, 국소적 처리를 위한 순환 신경망 층, 유효 변수 추출을 위한 변수 선택(Variable selection) 층, 유효하지 않은 특징 억제를 위한 개폐(Gating) 층을 사용한다[7]. 기존 트랜스포머와 달리 헤드 사이의 어텐션 가중치 값을 공유하여 추출한 특성의 중요도를 기반으로 어텐션 가중치를 계산함으로써 모델의 설명 가능성을 제고하였다.

본 연구는 Python 3.9.19 버전, pytorch-forecasting 1.0.0 버전 라이브러리를 사용하여 TFT 모델을 구현하였으며 전체 흐름도는 그림 1과 같다.



(그림 1) 전체 흐름도

2.2 데이터 수집

S&P500 주가지수는 상장지수펀드인 SPY 대응치로 사용하였다. 그 외에 주가 예측력을 가진 금융 데이터인 기간 스프레드, 하이일드 스프레드, VIX지수와 현재 금융 상황과 경기 전망을 반영하는 미국 10년물 국채수익률 데이터를 사용하였다. 데이터 수집은 yahoo finance에서 S&P500 지수와 VIX지수 데이터를, Federal Reserve Economic Data(Fred)에서 기간 스프레드, 하이일드 스프레드, 미국 10년물 국채 수익률 데이터를 각각 yfinance와 pandas_datareader 라이브러리를 통해 수집하였다. 기간은 2003년 1월 19일부터 2024년 7월7일까지 주별로 수집하였다.

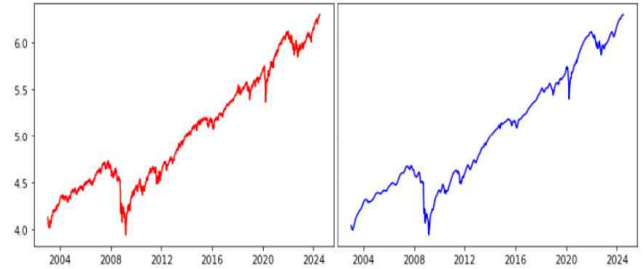
<표 1> 수집 데이터

변수명	설명
S&P500	S&P500을 1배로 추종하는 상장지수펀드
Term_Spread	미국10년물국채수익률 - 3년물국채수익률
High_Yield_Spread	투기등급 회사채 수익률 - 국채수익률
VIX	향후 30일간의 S&P500 지수의 예상 변동성 지수
Ten_year_treasury	미국 10년물 국채수익률

2.3 데이터 전처리

S&P500 데이터에 대해 로그 변환을 수행하였고 모든 변수에 대해 웨이블릿 변환으로 노이즈를 제거하였다. 웨이블릿 변환은 신호의 특징을 시간 및 주파수 도메인에서 동시에 분석하는 방법이다[8]. 주파수의 조율을 통해 추출한 상관관계를 분석하여 데이터의 잡음을 제거하는데 활용된다. 웨이블릿 함수는 선행연구에

따라 주가지수에 대해 가장 높은 SNR(신호 대 잡음 비)과 낮은 RMSE(평균 제곱근 오차)를 보이는 Coif3를 선택하였다[9]. 분계점 설정 방식은 계산이 간단하며, 일반적으로 적용할 수 있는 Universal threshold 방식을 사용하였다. 분계 처리 함수는 큰 계수에 대한 편향이 작고 노이즈 제거와 신호의 균형이 잡힌 Garrote 분계 함수를 선택하였다[10]. S&P500의 웨이블릿 변환 결과는 그림 2와 같다.



(그림 2) S&P500 노이즈제거 전(좌) 후(우)

이후, S&P500을 제외한 변수에 대하여 표준화를 수행하였다. 파생 변수로 변환된 S&P500에 이전 시점값을 차감한 S&P500_diff 변수를 생성함으로써 주가가 한 방향으로 계속 이동하려는 성질인 모멘텀 현상을 반영하고자 하였다. 상대적 시간 순서를 나타내는 변수인 time_idx와 연중 월과 주에 해당하는 month와 week 변수를 생성하였다. 최종적인 변수는 표2와 같다. 전처리된 데이터는 학습, 검증, 테스트데이터로 5 : 2.5 : 2.5 비율로 분할하였다.

<표 2> 최종 변수

종속변수	독립변수		
	미래시점에 알려지지 않은 변수	미래시점에 알려진 변수	etc
S&P500	S&P500	month	time_idx
	S&P500_diff		
	Term_Spread		
	High_Yield_Spread	week	
	VIX		
Ten_year_treasury			

2.4 모델링 및 평가, 시각화

모델은 52주(약 1년)의 과거 데이터를 통해 4주간의 미래를 예측하도록 설계하였으며 모델의 파라미터는 표 3과 같다.

모델의 평가 지표는 MAE, MAPE, SMAPE를 사용하였다. MAE는 예측값과 실제값 간 오차의 절댓값을 평균 낸 지표이며, MAPE는 예측값과 실제값 사이 오

차를 백분율로 변환 후 평균을 구한 지표, SMAPE는 예측값과 실제값 사이의 절대 오차를 두 값의 합으로 나누어 계산한 평가 지표이다.

<표 3> 모델 파라미터

파라미터	값
배치사이즈	64
LSTM 레이어 수	1
손실 함수	분위수 손실
에포크	100
그래디언트 클리프	0.1
학습률	0.01
히든 사이즈	512
어텐션 헤드 크기	4
드롭아웃	0.2

각 평가 지표의 계산 방법은 아래의 식 1, 2, 3과 같다. N은 데이터의 개수, A_i 는 실제값, P_i 는 예측값을 의미한다.

$$MAE = \frac{1}{N} \sum_{i=1}^N |A_i - P_i| \quad (\text{식 1})$$

$$MAPE(\%) = \frac{1}{N} \sum_{i=1}^N \frac{|A_i - P_i|}{|A_i|} \times 100 \quad (\text{식 2})$$

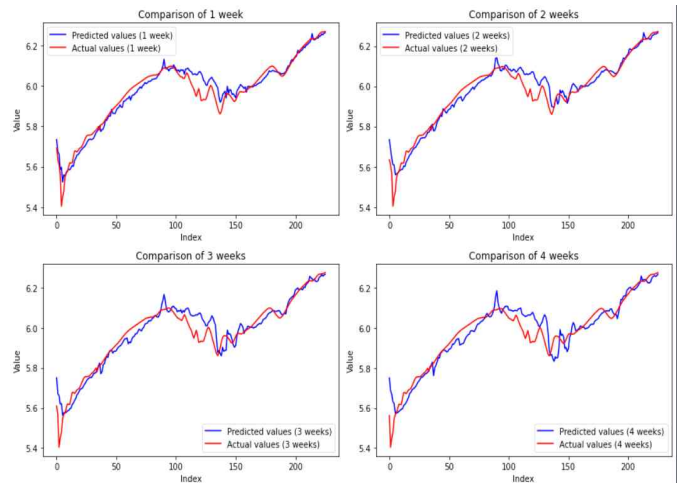
$$SMAPE(\%) = \frac{1}{N} \sum_{i=1}^N \frac{|A_i - P_i|}{(|A_i| + |P_i|)/2} \times 100 \quad (\text{식 3})$$

세부 수치는 표 4와 같다.

<표 4> 모델 평가 지표

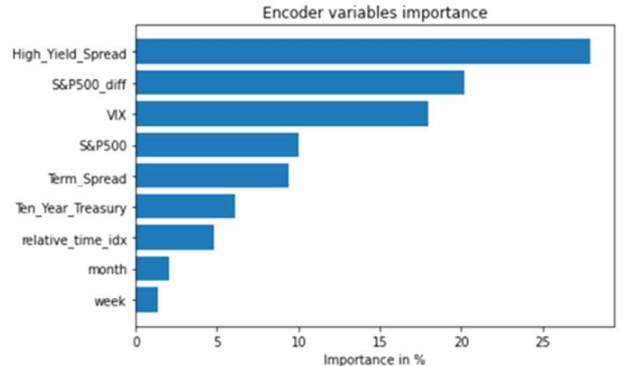
평가 지표	값
MAE	0.0329
MAPE	0.5559%
SMAPE	0.5541%

평가 지표를 이용한 정량적 평가 외에도 시각화를 통한 정성적 평가도 가능하다. 그림 3은 예측값과 실제값을 각 주차 별 예측 시점에 해당하는 예측값과 실제값을 비교한 그래프이다. 파란 실선은 예측 S&P500 지수를 의미하며 빨간색 실선은 실제 S&P500 지수를 의미한다. TFT 모델의 예측값 그래프와 실제값 그래프의 기울기가 유사함을 확인할 수 있다. TFT 모델은 예측을 할 때 과거 특정 시점에 대한 일반적 집중 정도와 변수의 중요도를 확인할 수 있어, 다각적인 결과 해석의 용이성을 높였다.

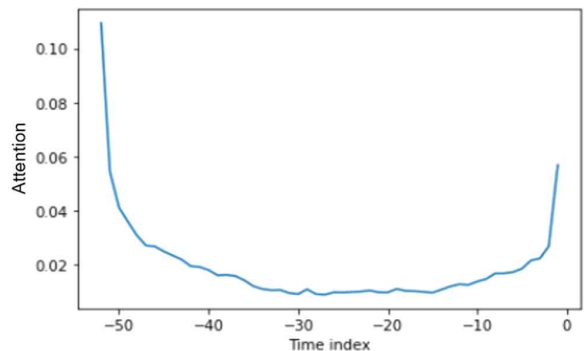


(그림 3) 예측값, 실제값 비교 그래프

그림 4는 변수의 중요도를 나타낸 그래프로 변수별 중요도를 파악하기 위해 9개 변수에 대한 시점별 변수 선택 가중치를 시간 축으로 평균낸 뒤, 그들의 합으로 나누어 나타낸 것이다[11]. 본 프로그램으로부터 목표 변수인 S&P500을 예측하기 위해 High_Yield_Spread, S&P500_diff 및 VIX 순으로 높은 중요도를 나타내었으며, 각각 약 28%, 약 20%, 약 18%의 수치를 나타내었다.



(그림 4) 변수 중요도



(그림 5) 시점 어텐션

그림 5는 시점 어텐션 그래프로 예측하는 첫 번째 시점(+0)에 대한 일반적 어텐션 패턴을 시각화한 것이다 [11]. 어텐션 헤드별로 평균을 취한 뒤 관측 기간 52 주의 기간 안에서 모든 변수에 대한 어텐션 가중치 합으로 계산된다. 즉, 변수와 관계없이 일반적으로 모델이 미래 시점을 예측할 때 어떤 시점에 집중하였는지 알아볼 수 있는 방법이다[11].

3. 결론

본 연구에서는 웨이블릿 변환을 통한 디노이징 기법과 설명 가능성이 높은 딥러닝 모델인 TFT 모델을 통하여 주가지수를 예측하였다. 예측 결과-테스트 데이터셋 간의 MAE, MAPE, 및 SMAPE는 각각 0.0329, 0.5559%, 0.5541%로 나타났다. 따라서 추후 연구에서는 장기적 예측을 시도하고, 기술적 지표와 경제변수 추가 및 정교한 모델링을 통하여 실무적 용이성을 제고하고자 한다.

참고문헌

[1] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen, "Machine learning explainability in finance: an application to default risk analysis.", Staff Working Paper No. 816, p.2, 2019

[2] Kelly, Bryan, and Dacheng Xiu, "Financial machine learning.", Foundations and Trends® in Finance 13.3-4, pp.205-363, 2023

[3] Samuelson, Paul A, "Rational theory of warrant pricing.", Henry P. McKean Jr. Selecta. Cham: Springer International Publishing, pp.195-232., 1965

[4] Fama, E. F., "Efficient Capital Markets: A Review of Theory and Empirical Work", The Journal of Finance 25(2), pp.383-417, 1970

[5] 이재웅, 한지형, "설명 가능한 KOSPI 증감 예측 딥러닝 모델을 위한 Layer-wise Relevance Propagation (LRP) 기반 기술적 지표 및 거시경제 지표 영향 분석.", 정보과학회논문지 48.12, pp.1289-1297, 2021

[6] Lim, Bryan, et al, "Temporal fusion transformers for interpretable multi-horizon time series forecasting.", International Journal of Forecasting 37.4, pp.1748-1764, 2021

[7] 김인경, 김대희, 이재구, "Temporal Fusion Transformers 와 심층 학습 방법을 사용한 다층 수평 시계열 데이터 분석." 정보처리학회논문지/소프트웨어

및 데이터 공학 제 11.2, p.2, 2022

[8] Chun-Lin, Liu, "A tutorial of the wavelet transform." NTUEE, Taiwan 21.22, p.2, 2010

[9] Qiu, Jiayu, Bin Wang, and Changjun Zhou, "Forecasting stock prices with long-short term memory neural network based on attention mechanism.", PloS one 15.1, e0227222, 2020

[10] Breiman, Leo. "Better subset regression using the nonnegative garrote." Technometrics 37.4, pp.373-384, 1995

[11] 한창진, "주가 방향성의 다중 시점 예측을 위한 해석 가능한 Transformer 모델 연구.", 서울대학교 대학원 석사학위 논문, pp.37-47, 2021